Overview

• Describing dirty data
  • Defining dirty data
  • Distinguishing different types of dirty data
  • Detrimental effects of dirty data

• Dealing with dirty data
  • Determining why respondents provide dirty data
  • Discouraging dirty data
  • Detecting dirty data
  • Deleting dirty data?
What is Dirty Data?

• Construct-irrelevant responses intentionally provided by study participants due to a lack of effort
  • Careless responding
  • Insufficient effort responding (IER)
  • Low-quality data

• Important differentiations:
  • Missing data
  • Outliers
  • Response sets
  • Social desirability
Types of Dirty Data

• Labels:
  • Random responding
  • Straightlining
  • Patterned responding
  • Satisficing

• Three classes of respondents\textsuperscript{1,2}:
  • Class 1: Normal responses (89% to 95%)
  • Class 2: “responding in an inconsistent way” (4% to 9%)
  • Class 3: “same response option for many consecutive items” (1% to 2%)

\textsuperscript{1}Meade & Craig (2012, quotes on p. 445), \textsuperscript{2}Maniaci & Rogge (2014)
Extreme Random Responding

• Random responding occurs when a participant’s response to one item has no bearing on his/her response to another item, regardless of the inter-item correlation

• Expected values$^3$:
  • Inter-item correlation: 0
  • Alpha: 0
  • First eigenvalue: 1

$^3$DeSimone et al. (2018)
• Straightlining occurs when a participant provides identical responses to consecutive items

• Expected values\(^3\):
  • Inter-item correlation: 1
  • Alpha: 1
  • First eigenvalue: J (#items)

\(^3\text{DeSimone et al. (2018)}\)
Complications in Real Data

• The expected inter-item correlation for a dataset comprising 100% consistent random responders and straightliners is equal to the proportion of the sample who are straightliners
  • Not all respondents provide dirty data (< 100%)
  • Not all respondents who provide dirty data do so consistently
  • Random responders can also straightline (and vice versa)
  • There may be other forms of dirty data

• Any amount of dirty data is undesirable if it has the potential to influence the results of psychometric or statistical analyses
Influence of Dirty Data on Psychometrics

• Simulated data\textsuperscript{3}: Proportion of dirty required to change the conclusions drawn from common psychometric estimates

<table>
<thead>
<tr>
<th></th>
<th>(\alpha)</th>
<th>(r_{ij}) (SRMR)</th>
<th>(\lambda) (CL)</th>
<th>Structure (PCA)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random responding</td>
<td>65%</td>
<td>30%</td>
<td>15%</td>
<td>85%</td>
</tr>
<tr>
<td>Straightlining</td>
<td>n/a</td>
<td>&lt;5%</td>
<td>10%</td>
<td>20%</td>
</tr>
</tbody>
</table>

• Real data: Effects of screening techniques on psychometric estimates
  • \(\alpha\) lower for inattentive respondents\textsuperscript{2}, though \(\alpha\) can increase or decrease depending on which type of dirty data is removed\textsuperscript{4}
  • Inter-item and inter-scale correlations change more when removing straightliners than when removing random responders\textsuperscript{4}

\textsuperscript{2}Maniaci & Rogge (2014), \textsuperscript{3}DeSimone et al. (2018), \textsuperscript{4}DeSimone & Harms (2018)
• Dirty data can increase or decrease correlations, depending on the type and distribution of dirty data\textsuperscript{2,4}
  • For example, if respondents providing dirty data are more likely to endorse responses near the midpoint of a scale, dirty data may inflate correlations\textsuperscript{5,6}

• Counterintuitively, removing participants who provide dirty data can increase statistical power\textsuperscript{2}

\textsuperscript{2}Maniaci & Rogge (2014), \textsuperscript{4}DeSimone & Harms (2018), \textsuperscript{5}Credé (2010), \textsuperscript{6}Huang et al. (2015)
Why Respondents Provide Dirty Data

• Traditional perspectives\textsuperscript{7}:
  • Distraction
  • Lack of interest
  • Laziness
  • Fatigue

• Recent empirical work on the nomological network of dirty data
  • Self-reports
  • Peer-reports, experiments and indirect measures

\textsuperscript{7}Krosnick (1991)
Self-report Correlates of Dirty Data

• Five-factor model traits and their facets\(^6\)
  • Neuroticism (positive)
    • Depression, vulnerability
  • Extraversion (negative)
    • Cheerfulness, friendliness
  • Openness (negative)
    • Artistic interests, imagination
  • Agreeableness (negative)
    • Altruism, trust
  • Conscientiousness (negative)
    • Self-efficacy, dutifulness

\(^6\)Huang et al. (2015)
Self-report Correlates of Dirty Data

• HEXACO traits
  • Honesty/Humility (negative)
  • Emotionality (negative)
  • Conscientiousness (negative)

• Dark traits
  • Machiavellianism (positive)
  • Psychopathy (positive)

• Academic outcomes
  • GPA (negative)
  • Class absences (positive)

---

8McKay et al. (2018), 9Bowling et al. (2016)
Other Correlates of Dirty Data

- Peer-reported personality\(^9,10\)
  - Neuroticism (positive)
  - Extraversion (negative)
  - Agreeableness (negative)
  - Conscientiousness (negative)

- Implicit aggression (positive)\(^10\)

- Questionnaire length (mixed results)\(^11\)

\(^9\)Bowling et al. (2016), \(^10\)DeSimone et al. (in press), \(^11\)Gibson & Bowling (in press)
Preventing Dirty Data

- Instructional warnings\(^1,11\)
- Promising rewards\(^11\)
- Asking participants nicely
- Creating a “living survey”

\(^{1}\)Meade & Craig (2012), \(^{11}\)Gibson & Bowling (in press)
If I think hard enough, I can often control the weather.

In the last block of items, you indicated that you could often control the weather. Your thoughtful and effortful responses to this survey are important for the advancement of research in our field. Please try to answer questions more carefully from this point forward.
How to Identify Dirty Data

- Data screening indices\(^{12,13}\)
  - Direct
    - Self-report (or “use me”) questions
    - Bogus or instructed items
  - Archival
    - Response time
    - Longstring/individual response variance (IRV)
    - Semantic synonyms/antonyms
  - Statistical
    - Psychometric synonyms/antonyms
    - Personal reliability
    - Mahalanobis \(D\)

\(^{12}\text{DeSimone et al. (2015), }^{13}\text{Curran (2016)}\)
## Screening Techniques

<table>
<thead>
<tr>
<th>Technique</th>
<th>Designed to Capture</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self-report</td>
<td>Confessions of dirty data providers</td>
</tr>
<tr>
<td>Bogus or Instructed Items</td>
<td>Random responding and straightlining</td>
</tr>
<tr>
<td>Response Time</td>
<td>Random responding and straightlining</td>
</tr>
<tr>
<td>Longstring</td>
<td>Straightlining</td>
</tr>
<tr>
<td>IRV</td>
<td>Straightlining</td>
</tr>
<tr>
<td>Synonyms and Antonyms</td>
<td>Random responding</td>
</tr>
<tr>
<td>Personal Reliability</td>
<td>Random responding</td>
</tr>
<tr>
<td>Mahalanobis’ $D$</td>
<td>Atypical responding</td>
</tr>
</tbody>
</table>
Data Screening Considerations

- Screening after the results are known ("SARKing")
- Cutoffs
- Missing data
- Reverse scoring
- Balanced scales
- Screening for multiple types of dirty data
- Maintaining the integrity of scales
- Respondent inconsistency
- Respondent frustration
What to do with Dirty Data?

• Arguments for removing
  • Dirty data is a source of construct-irrelevant variance\(^4\)
  • Dirty data can reduce statistical power\(^2\)

• Arguments for retaining
  • False positives
  • Relationships with focal constructs\(^{10}\)

• Make an *a priori* choice based on your goal

\(^2\)Maniaci & Rogge (2014), \(^4\)DeSimone & Harms (2018), \(^{10}\)DeSimone et al. (in press)
Recommendation 1: Anticipate Dirty Data

• Incorporate preventative measures and data screening into study design

• Consider which type(s) of dirty data are the most likely, interesting, or problematic

• Determine screening techniques and cutoffs before beginning data collection

• Plan to identify (and possibly remove) dirty data when considering target sample size
Recommendation 2: Screening Techniques

• Use multiple techniques designed to capture multiple types of dirty data

• Always time participant responses, and always consider using other unobtrusive screening techniques

• Use the correct screening techniques for your research design
  • Item wording and direction (balance)
  • Scale length
  • Respondent perception

• Screen conservatively, but scrutinize data
Recommendation 3: Consider Using Dirty Data

- Providing dirty data can be considered an observable behavior.
- The tendency to provide dirty data has been demonstrated to correlate with personal characteristics and situational factors.
- Dirty data may influence the focal relationships in a study.
- Consider modeling or controlling for dirty data as opposed to discarding dirty data.
Recommendation 4: Transparency

• Report the details of screening for dirty data
  • Techniques used
  • Cutoffs
  • Deviations from the original plan
  • Determinations of whether or not to remove dirty data
  • Consider sensitivity analysis

• Report results before and after screening
  • How many respondents were flagged for providing dirty data
  • Was dirty data related to any focal constructs
  • How did removing dirty data affect the study results


